June 12, 2023

National Telecommunications & Information Administration
U.S. Department of Commerce
1401 Constitution Ave., N.W.
Washington, D.C. 20230

Re: NTIA AI Accountability Policy Request for Comment(Docket number 230407-0093)

The Integrity Institute appreciates the opportunity to provide input to NTIA's "AI Accountability Policy Request for Comment." As AI systems are being rapidly deployed in public, it is crucial that society understands our inability to walk back the real-life implications of building and deploying these technologies irresponsibly. As Integrity Workers, we believe we can bring our expertise to bear on this moment and help foster the development of responsible AI. We respectfully offer the comments below in response to the questions posed, and hope to remain engaged in the efforts in this area.

Sincerely,

Sahar Massachi
Co-Founder and Executive Director
Integrity Institute

Jeff Allen
Co-Founder and Executive Director
Integrity Institute

Integrity Institute Comments in Response to NTIA AI Accountability Policies Request for Comment (Docket number 230407-0093)

Introduction

The Integrity Institute is a 501(c)(3) think tank powered by a community of integrity professionals: tech workers with experience in integrity roles — roles dedicated to fixing harms to people and society within social internet platforms.

Institute members have observed, and many have helped build, the architecture of the social internet. They have direct experience tackling spam, hoaxes, harassment, hate speech, disinformation and more. They understand the systemic causes of problems on the social internet and how to mitigate or avoid them. They have seen (and built!) successful and unsuccessful solutions, and can tell the difference. The Integrity Institute is committed to sharing member expertise directly to the people theorizing, building, and governing the social internet, so that the social internet can help individuals, societies, and democracies thrive.

While the focus of our work has been social media platforms, we are paying attention to these conversations around AI for two significant reasons: first, we are seeing AI developments (and generative AI in particular) have impacts on the work of Integrity professionals. Institute members have recently published *a blog post* about the ways that AI can present opportunities, challenges and solutions for those people working on trust and safety online.

Second, we believe that we can bring our expertise to bear in this conversation in a constructive way. The experience of Integrity Workers in tackling complex problems of the social internet, understanding risks, the role that design plays in amplifying or mitigating those risks, thinking through unknown impacts and scenarios and the tough choices that come with governing these spaces, is directly applicable to the current discussions around building responsible AI and designing effective accountability mechanisms. There is a parallel in the current moment with the arc of social media platforms: rush to market, safety was not designed in, and once products were launched the companies (and world) saw new ways it could be used in negative ways. Integrity Workers are the ones trying to combat these problems and encourage responsible design. There is an opportunity here with AI to think through these things from the beginning, build in safety, fairness and transparency. Accountable AI policies from NTIA would be a start.

Comments prepared from the following resources:

1. Recent blog post from II members outlines some basic ways Integrity Workers can enter the conversation on responsible AI.: *https://integrityinstitute.org/blog/why-ai-may-make-integrity-jobs-harder*
2. Blog post on how AI will impact the work of Integrity Professionals: *https://integrityinstitute.org/blog/unleashing-the-potential-of-generative-ai-in-integri[…]st-amp-safety-work-opportunities-challenges-and-solutions*
3. II Feedback on the EU AI Act, including a segment on AI audits: *https://integrityinstitute.org/blog/comment-on-eu-ai-act*
4. Extensive inputs from David Harris, Integrity Institute Member

Responses to Questions

AI Accountability Objectives

1. What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?

a. What kinds of topics should AI accountability mechanisms cover? How should they be scoped?

Accountability mechanisms on the whole should contribute to transparency and explainability of the AI system and its use, e.g., demonstrate how individuals affected by an AI can see and understand how such decisions were made. They should cover the harms and risks of harms to individuals as well as to societies where the AI system operates. Where harms could include exposure to illegal and harmful content, bias and exclusion, and violation of privacy, security, and human rights. And the accountability mechanisms should cover the scale of harms, cause of harms, and nature of the harms.

Requirements for AI audits (or other accountability mechanisms) should include transparency requirements for auditors and users (or those impacted by an AI system). For example, audit requirements should stipulate that auditors have access to raw data and can reproduce findings from model/system developers, rather than just trusting the developers to provide conclusions about their systems. In general, accountability mechanisms

In the view of some of our members, AI accountability mechanisms should clearly demonstrate that the producers of AI systems understand exactly how their systems are working and have fully studied the system design and outcomes and are willing to take responsibility for them in clear ways. They should require public reporting in the same way that other industries like health care and

construction have detailed codes and practices to make sure that organizations comply with best practices.

Audits in particular should be scoped broadly, in a way that allows for fundamental questions to be asked beyond whether an AI system is exhibiting bias – including, for example, if the AI system should exist at all. Even for audits limited to assessing bias, the traditional approach of employing model cards would not be sufficient. A more system-wide approach would be needed to truly understand the impacts of the AI use.

Other components of various mechanisms could include red teaming exercises to stress test high risk systems in various scenarios (testing for bias, among other impacts). Any audit or assessment should include disclosure of how data collected from people is used and how the system makes predictions of what people will do. This is particularly cogent for AI systems that are used to recommend content to people (e.g., on social platforms), and such data disclosure should give transparency into how the system responds to harmful content, which can reveal any relationship between the prevalence of harmful content and the system design.

b. What are assessments or internal audits most useful for? What are external assessments or audits most useful for?

Internal assessments are useful in making sure that organizations, especially large organizations, are giving employees concerned with the safety of the system influence in its design, maintenance, and operation and have a clear handle on what their AI systems are doing. Like other system audits, they are also useful for assessing to what extent the system is contributing to the values and mission of the organization. Organizations should be developing systems in line with their values and corresponding metrics to measure against that. Internal assessments are useful for testing changes to the system or different applications of the system, and understanding implications and associated risks. Such insights can (should) lead to the development of policies or technical measures to mitigate risks.

External assessments and audits are useful for holding organizations accountable for the safety and fairness of their AI systems, which is crucial for achieving public trust. External assessments should incentivize companies to build more responsibly, and give companies the opportunity to demonstrate their responsibility when they are. It is also important that users and those impacted by the AI systems understand the role the system played in making decisions relevant to them, and have insight into the characteristics and evaluation of the system.

c. An audit or assessment may be used to verify a claim, verify compliance with legal standards, or assure compliance with non-binding trustworthy AI goals. Do these differences impact how audits or assessments are structured, credentialed, or communicated?

Some members view these differences as having significant impacts, specifically in that there are legal consequences for failing the first two types. Such mechanisms should be rigorous, certified to meet existing industry standards (meaning the industry in which the AI system is deployed, e.g, legal, health, etc.), involve third parties as a means to combat any bias, and mandatory for any organization deploying an AI system in these contexts. The requirement could be tied to other industry certifications. The third (assuring compliance with non-binding trustworthy AI goals) presents opportunities for gaming the system, as mechanisms may not be as rigorous or involve the same compliance requirements, however, they should still require disclosure of data and information to third parties to strengthen accountability and avoid bias in reporting results.

d. **Should AI audits or assessments be folded into other accountability mechanisms that focus on such goals as human rights, privacy protection, security, and diversity, equity, inclusion, and access? Are there benchmarks for these other accountability mechanisms that should inform AI accountability measures?**

Some of our members believe they should not be folded into other mechanisms, but they should adopt and/or adapt the principles, benchmarks and legal standards associated with those goals. Assessments and audits should be designed to create accountability towards those goals and give the public transparency around how the AI systems are performing against them. However, benchmarks need significant work to be meaningfully adapted to AI systems. This is an area for further study.

e. **Can AI accountability practices have meaningful impact in the absence of legal standards and enforceable risk thresholds? What is the role for courts, legislatures, and rulemaking bodies?**

Comprehensive transparency around the scale, cause, and nature of harms that an AI system is causing, or the risk of which is it creating, would, on its own, create a significant incentive to build more responsibly. Voluntary standards and accountability practices can be useful in shaping and incentivizing behavior over the long term. However, given the speed at which these systems are being deployed for use by the general public, there may be a need for legal standards and enforceable risk thresholds, otherwise the incentives for developers will continue to be to create and deploy systems as fast as possible without adequate assessment and consideration of the risks and impacts.

Comprehensive transparency is not something that we should expect companies to provide on their own, however. Some of our members take the position that accountability practices cannot have meaningful impact without legal standards and enforceable risk thresholds, as otherwise, many, if not most, developers of AI systems will produce dangerous and discriminatory AI systems that do not serve the public interest. Courts, legislatures and rulemaking bodies must urgently develop

competency in AI systems and block the usage and/or release of AI systems that have not yet been thoroughly assessed for risk. This is very similar to the way that the FDA does not allow drugs to be sold without careful study.

2. Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

Both of these angles are important, and policies can address both by building in requirements for transparency and external audits. Transparency will contribute to trust for external stakeholders, while also creating incentives and mechanisms for holding organizations accountable for developing AI in a safe, fair and responsible manner.

In the view of some of our members, the *greatest* value for such mechanisms is internal. That is, certifications, audits and assessments should be designed by policymakers to transform the internal processes of organizations that develop AI systems. They should incentivize the developers to put safety and fairness front and center and transparently demonstrate that they are doing this in a continuous manner. They should upend the current race-to-the bottom model of releasing AI systems as quickly as possible without serious work to mitigate potential consequences.

3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?

a. **The AI system does not substantially contribute to harmful discrimination against people.**
b. **The AI system does not substantially contribute to harmful misinformation, disinformation, and other forms of distortion and content-related harms.**
c. **The AI system protects privacy.**
d. **The AI system is legal, safe, and effective.**
e. **There has been adequate transparency and explanation to affected people about the uses, capabilities, and limitations of the AI system.**
f. **There are adequate human alternatives, consideration, and fallbacks in place throughout the AI system lifecycle.**
g. **There has been adequate consultation with, and there are adequate means of contestation and redress for, individuals affected by AI system outputs.**
h. **There is adequate management within the entity deploying the AI system such that there are clear lines of responsibility and appropriate skillsets.**

These goals present tradeoffs, but they are by no means insurmountable. Tensions between privacy and transparency, for example, are well-documented in the context of social media platforms, but we

believe that meaningful transparency–in most cases–can be achieved without violating individual privacy rights.

The view of some Institute members is that assessing these measures will require significant expertise from whomever is responsible—any team would need to have people with backgrounds in AI technical development, data science, social science, corporate governance, law and policy. This is possible for larger organizations to set up in house, and will be costly. A single instrument could be used, but it would be highly complex and likely take months for an internal or external team to assess, and if an external team is responsible for it, that team would need unfettered internal access and significant training.

4. Can AI accountability mechanisms effectively deal with systemic and/or collective risks of harm, for example, with respect to worker and workplace health and safety, the health and safety of marginalized communities, the democratic process, human autonomy, or emergent risks?

Systemic approaches to audits or assessments can be undertaken, which would address these systemic risks. These approaches look beyond just evaluating the AI model for certain attributes (e.g., bias) and take a holistic approach.

Some II members noted that not only can mechanisms deal with these risks, but all developers of AI systems should be required to demonstrate that their AI systems have been thoroughly assessed for these types of risks and that all efforts possible have been taken to reduce or mitigate these risks. If the risks cannot be fully mitigated, the developer must demonstrate that the value to society is so immense that the risks are worthwhile. This is similar to the way in which environmental impact assessments are conducted before large construction projects, and if the impact is deemed too large, the project cannot be approved.

5. Given the likely integration of generative AI tools such as large language models (*e.g.,* ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI? [80]

Some of our members take the position that if the tools do not comply with standards for trustworthy AI, they should not be launched. Developers should be legally required to notify any people who use their products that AI systems were used in the production of those products, and that those products may have risks or problems associated with them. However, this is an area for further evaluation and study, where Integrity Workers could have valuable insights.

7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

Some of our members take the firm stance that accountability mechanisms are absolutely critical to technologies that pose such significant threats to society. However, regulation around implementation details of AI, rather than regulation around transparency of harms caused, could end up frustrating the development of safer AI. Requiring that systems are built in a certain way could be detrimental if a safer way was developed.

In addition, some members have expressed the view that industry should be incentivized to pursue a "Responsible AI Race." Integrity workers can contribute to this by partnering with organizations teaching product development best practices or project management strategies to incorporate privacy, bias, addiction and data use/user education relating to AI and underlying algorithms. We can reach beyond the typical channels and we as integrity workers can partner with employers and provide guidance on employee resources for companies, such as recommended language to incorporate into Employee Handbooks. Across the industry, it will be critical to ensure proprietary information, trade secrets or personally identifiable information (PII) is not divulged for the purposes of increased productivity on a task.

Barriers to Effective Accountability

24. What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?

Some members have identified the significant challenge in the current AI context is the race to market happening among industry players. Instead of the norm of deploying features *after* testing, the current breakneck pace of development has led to a system of “post-hoc safety” where users of the product are providing the testing in production. There is a lack of incentive for companies to build AI systems responsibly from the start, and public policy that mandates transparency, assessments, and audits would be beneficial.

Without significant consequences to provide enforcement, accountability mechanisms will be seen as an avoidable or gameable nuisance for companies intent on deploying products quickly. To change the incentives, there is a role for Integrity Workers to contribute to creating demand-side pressure in the market to develop responsibly and fulfill accountability requirements demonstrating that. Some II members have articulated specific ways they see that Integrity Workers can play this role:

- Help **create AI literacy** - both by influencing tech companies' policies and operations; by working with regulators and local policy makers and through our internal, external, and *exterior* partnerships (that is, those occurring outside of a role or not contained to a position with a company, which further engages in the industry community through participation in think tanks, like the Integrity Institute).
- Openly **discuss and present concerns** related to AI, such as privacy, bias, exposure to harms, and the potential for automation to displace the *'human in the loop'* (and resultant dangers of this). We can discuss the specific threat types as we understand them today and create a better information sharing environment for the future. Integrity workers can participate in different forums to hold these discussions. Through podcasts, conferences and local integrity meetings/virtual training, or even through social media posts and broadcasts, we can help people understand the broader implications of use while we actively assess the evolving risks of using AI. This includes educating users on the known-unknowns while working with AI developers/organizations on addressing the unknown-unknowns, and how they can incorporate these into transparency reports.
- **Create a trusted global workforce in Responsible AI**. We are all faced with the consequences of how GenAI models are built and deployed regardless of industry or occupation. The way products are coming to market in the AI space can be dangerous if we continue this race to market product deployment approach. Instead of an "AI Arms Race" we should pursue a "Responsible AI Race." We can do so by partnering with organizations teaching product development best practices or project management strategies to incorporate privacy, bias, addiction and data use/user education relating to AI and underlying algorithms. We can reach beyond the typical channels and we as integrity workers can partner with employers and provide guidance on employee resources for companies, such as recommended language to incorporate into Employee Handbooks. Across the industry, it will be critical to ensure proprietary information, trade secrets or personally identifiable information (PII) is not divulged for the purposes of increased productivity on a task.
- We, as integrity workers, **can supply general considerations** for policymakers, regulators and other organizations, outlining the risks that might accompany the use of AI products to perform quick analysis or tasks without a consideration of failure modes. But executives will have to invest in the Integrity workforce, including AI-related training and upskilling, not just add to their current T&S and content moderation workload. Companies must do more than just outsource the work of training and moderating LLMs to underdeveloped countries with abundant cheap labor.
- **We can cross-pollinate the industry with experts to train people to help create a safer internet**. Just as we recommend social internet companies prepare for civic operations as best as they can by predicting the outcome of societal reactions before, during and after a civic event (like an election), we can apply this practice in the context of AI. We should learn from other industries who have gone through exponential growth (think the invention and adoption of the rotary action steam engine during the industrial revolution). Or, we can borrow models from organizations such as the Digital Forensics Lab operated by The Atlantic

Council Digital Sherlock Program (designed to train people to detect information manipulation on the social internet using OSINT techniques). These examples offer creative insights and practices to building a safer social internet in this new context.